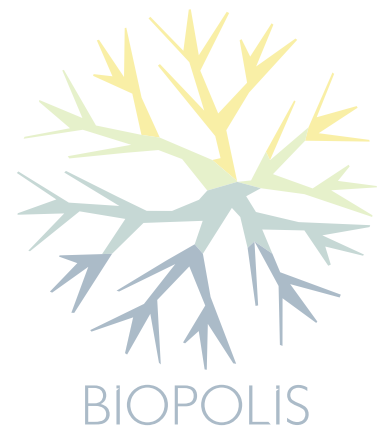


WP4

Revision of the Data Management Plan Including ORPD: Open Research Data Pilot

Deliverable 4.8



Revision of the Data Management Plan Including ORPD: Open Research Data Pilot

Deliverable 4.8

Lead beneficiary	BIOPOLIS
Submission date	April 5 th 2024



BIOPOLIS

Deliverable 4.8 (D4.8)

Revision of the Data Management Plan Including OPRD: Open Research Data Pilot

Project acronym: BIOPOLIS
Project name: Teaming to Upgrade to Excellence in Environmental Biology
Ecosystem Research and Agrobiodiversity Work
Programme Topics Addressed: H2020-WIDESPREAD-2018-01 (TEAMING)
Grant agreement: 857251
Project duration: 01/10/2019 – 30/09/2026 (84 months)
Coordinator: Associação BIOPOLIS

Delivery date of first version: M48 (March 2024)
1
Lead beneficiary: BIOPOLIS
Project's coordinator: Nuno Ferrand

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	



TABLE OF CONTENTS

Summary.....	3
1. Introduction.....	4
2. Data Management and Open Access Committee (DM&OAC).....	5
2.1. Mission and Objectives	5
2.2. Composition	6
3. Data Management Plan (version 2).....	7
3.1. Data Types.....	8
3.2. Data Management - Handling, Storage, and Archiving	9
3.2.1. Nucleic Acid Sequence Data	9
3.2.2. Taxonomic Data.....	10
3.2.3. Molecular Biological or Field Methods.....	11
3.2.4. Ecological Results Dissemination	12
3.2.5. Camera Trap Images.....	13
3.2.6. GPS-tagged Animal Data	14
3.2.7. Vegetation Data	14
3.2.8. Land Use and Land Cover Data.....	15
3.2.9. Conventional Biodiversity Survey Data	15
3.3. Data Analysis Methods	16
3.4. Data Dissemination - Sharing and Publishing.....	16
3.5. Software.....	17
3.6. DMP Implementation	18
3.7. Data Security, Research Ethics, and Animal Welfare.....	18
4. DMP Implementation Actions and Status.....	19
4.1. DMP Implementation Status	19
4.2. DMP Implementation Actions.....	21
4.3. DMP Implementation KPIs.....	22
5. Concluding Remarks	23

Summary

This document presents a revised version of the BIOPOLIS Data Management Plan (DMP), originally submitted as Deliverable D4.7 in March 2020 (Month 6), which presents the main elements of the data management policy used by the BIOPOLIS Teaming project, funded by the European Union's H2020 Programme under Grant Agreement No.857251.

Key Updates

- *Version 2 of the BIOPOLIS DMP*: Included in Section 4, this revised DMP details the types of data collected, data management practices, and data sharing protocols aligned with Horizon EUROPE's Open Research Data principles and broader open science practices. The DMP will be further updated during the project's lifetime if deemed necessary by the newly established Data Management and Open Access Committee (DM&OAC).
- *Establishment of the DM&OAC*: The DM&OAC was recently established to ensure effective implementation of the DMP. The DM&OAC is responsible for overseeing the DMP's implementation and further refinement within the next year. This document also outlines the DM&OAC's mission, objectives, and composition. The DM&OAC selection process is ongoing and should be concluded by the end of April 2024.
- *Data Management Implementation Update*: Section 5 provides a progress report on DMP implementation, outlining key actions planned and the development of Key Performance Indicators (KPIs) for data management. Efforts are underway to establish methods for collecting data necessary to calculate these KPIs.

1. Introduction

The European Commission (EC) funded BIOPOLIS project is upgrading CIBIO - Research Centre in Biodiversity and Genetic Resources, into a leading Centre of Excellence (CoE) for Environmental Biology, Ecosystem Research, and Agrobiodiversity. This is achieved through close collaboration (Teaming) with the renowned University of Montpellier (UM), France, and Porto Business School (PBS), Portugal.

BIOPOLIS focuses on three key areas:

- i) Environmental and Biodiversity Assessment and Monitoring: This area develops methods to evaluate and track environmental health and biodiversity.
- ii) Ecosystem Function and Restoration: This area concentrates on understanding ecosystem functioning and developing strategies for restoration.
- iii) Agrobiodiversity, Conservation, and Competitiveness: This area addresses the preservation of local genetic resources and farming systems, along with enhancing their competitiveness.

This document presents a revision to the BIOPOLIS Data Management Plan (DMP), submitted as Deliverable D4.7 in March 2020 (M6), and subsequently approved. It is expected that the research activities will generate a significant amount of diverse data that will require careful management and sharing. To ensure effective implementation of this DMP, the Data Management and Open Access Committee (DM&OAC) was recently established. The DM&OAC is responsible for overseeing the DMP's implementation and further refinement within the next years. This document also outlines the DM&OAC's mission, objectives, and composition. Key performance indicators (KPIs) for data management are here defined and work is underway to develop methods for collecting data necessary to calculate these KPIs.

2. Data Management and Open Access Committee (DM&OAC)

2.1. Mission and Objectives

The DM&OAC is dedicated to cultivating a research environment at BIOPOLIS-CIBIO that prioritises responsible data management and open-access practices. This commitment aligns with European Union¹ and Portuguese² open science policies and aims at fostering research reproducibility, transparency, and the advancement of knowledge for the benefit of the wider scientific community and public.

The main objectives are:

- **Develop and implement comprehensive policies and procedures** for data management and open access, ensuring alignment with national and international standards.
- **Evaluate and approve research data management plans** submitted by researchers, ensuring compliance with institutional and ethical guidelines.
- **Monitor adherence to data management policies and procedures**, proactively addressing any identified issues to maintain data integrity and security.
- **Champion open access to research data** while balancing ethical and legal considerations, promoting public access to valuable research findings.
- **Raise awareness** among researchers about the significance of responsible data management and open access, fostering a culture of data stewardship and transparency within the institute.

¹ EU Open Science: https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en

² Portuguese Open Science: <https://www.fct.pt/en/sobre/politicas-e-estrategias/politicas-de-ciencia-aberta/>

2.2. Composition

The DM&OAC will be a **diverse and inclusive body** comprised of individuals with expertise in:

- Data management
- Open access
- Ethics and law
- Research disciplines
- Information technology and infrastructure

Membership includes:

- Researchers representing various disciplines
- Legal or ethics expert
- IT or infrastructure representative

The committee will have a **chair or co-chairs** to lead and coordinate its activities. The committee will have **3 or 5 members**, with the final number determined by the expertise needed to fulfil the objectives and ensure diverse representation. Committee members will serve a two-year term, with the possibility of extension for up to three terms in total.

The committee will be led by either a chair or co-chairs, chosen by the Board of Directors based on their expertise and leadership skills. The final number of members, either 3 or 5, will be determined by the specific expertise needed to fulfil the objectives and ensure diverse representation across research disciplines, data management, legal/ethics, and IT.

Selection Procedure

Committee members will be selected through a fair and open process (transparent and inclusive) to serve a two year mandate. People from the BIOPOLIS-CIBIO research community and key stakeholders (people or groups with a vested interest) may apply. The process aims to find people with i) needed knowledge and skills (expertise); ii) Strong dedication to the committee's goals (commitment); and iii) A wide range of viewpoints (diverse perspectives).

The executive committee will review all applications based on the following three main criteria while trying to create a balanced committee, including gender balance, with a variety of backgrounds (balanced representation) and ensuring the committee can function effectively (effective functioning of the DM&OAC):

- Relevant knowledge and experience (expertise)
- Past work experience (experience)
- Support for responsible data management, open data sharing, and FAIR principles (commitment to responsible data management and open data and FAIR principles).

The DM&OAC selection process is ongoing and should be concluded by the end of April 2024.

3. Data Management Plan (version 2)

The BIOPOLIS Data Management Plan (DMP) outlines how the project handles research data to ensure it adheres to the FAIR principles (Findable, Accessible, Interoperable, Reusable) as outlined in the Guidelines on FAIR Data Management in Horizon 2020 and aligns with the Horizon Europe Open Science Policy. This approach guarantees data discoverability, and accessibility, while promoting its reuse for future research, ultimately enhancing data quality and reproducibility.

The DMP leverages expertise gained from CIBIO's previous EC-funded projects, namely:

- ERA Chair in Environmental Metagenomics (EnvMetaGen) (Project number 668981):
This project provided valuable knowledge and best practices for data management.
- ERA Chair project in Tropical Biodiversity – TROPiBIO (Project number 854248):
Synergies with TROPiBIO's DMP further strengthen BIOPOLIS' data management strategy.

Additionally, the DMP aligns with:

- PORBIOTA – Portuguese e-Infrastructure for Information and Research on Biodiversity (<http://www.porbiota.pt/en>): Coordinated by CIBIO, PORBIOTA provides the essential infrastructure for biodiversity data management.
- LifeWatch ERIC (<https://www.lifewatch.eu/>): As the Portuguese node of LifeWatch ERIC, BIOPOLIS contributes to a broader European effort for biodiversity data management.

This section includes the second version (version 2) of the BIOPOLIS DMP. The newly established DM&OAC (Data Management and Open Access Committee) will oversee and update this DMP throughout the project's lifecycle and beyond, as needed.

3.1. Data Types

The BIOPOLIS project generates a heterogeneous set of datasets encompassing various aspects of biodiversity and ecosystems, including:

- Georeferenced Biodiversity: Data on species location, distribution, and abundance.
- Ecosystem Composition, Health and Services: Information on the composition, health and functionality of ecosystems and their services.
- Genomic Data: Data obtained through next-generation and other sequencing technologies.

The main data types include but are not limited to:

- Nucleic acid sequence data (including e.g. genomic breeding values);
- Taxonomic information and associated metadata;
- Molecular biological and field collection methods;
- Ecological information based on the results of experimental work;
- Camera trapping and GPS tag data;
- Vegetation data;
- Land use and land cover data (including socio-ecological information);



- Conventional biodiversity survey data (species presence/absence and abundance);
- Data regarding analysis methods (including computer modelling and simulation);

Each data type will be archived and disseminated following appropriate methods based on its specific characteristics. Links to the repositories holding these datasets will be centralised in the BioStudies³ database organised by project/study. Specific data types may also be archived in additional specialised databases depending on their nature. Details on these additional repositories will be provided in the following sections.

3.2. Data Management - Handling, Storage, and Archiving

3.2.1. Nucleic Acid Sequence Data

Nucleic acid sequence data archiving strategies vary based on the extent of analysis and annotation of features completed on the data.

Raw sequence reads and associated sample metadata should be deposited in the European Nucleotide Archive (ENA)⁴. The ENA is part of the ELIXIR Core Data Resources⁵, ensuring open access and long-term preservation.

Sequenced data with confident taxonomic assignment (e.g., species level or below) by a recognised taxonomic expert should continue to be deposited, whenever possible, in additional relevant databases beyond the primary data repository:

- GenBank⁶: A comprehensive public database of nucleotide sequences maintained by the National Center for Biotechnology Information (NCBI).

³ BioStudies database: <https://www.ebi.ac.uk/biostudies/>

⁴ European Nucleotide Archive (ENA): <https://www.ebi.ac.uk/ena/>

⁵ ELIXIR Core Data Resources: <https://elixir-europe.org/>

⁶ GenBank: <https://www.ncbi.nlm.nih.gov/genbank/>

- Barcode of Life Database (BOLD)⁷: An open-access repository specifically focused on DNA barcodes for species identification.

These chosen archives prioritise data discoverability, open accessibility, reproducibility, and reusability. Deposited data will be readily available for future analysis and reinterpretation by the scientific community. Standardised protocols will be implemented to ensure consistent and, whenever possible, semi-automated submission of sequence data to the aforementioned archives.

3.2.2. Taxonomic Data

This section outlines the preferred archiving strategy for taxonomic information associated with research projects funded by BIOPOLIS, emphasising FAIR (Findable, Accessible, Interoperable, Reusable) principles.

As mentioned above, sequences that can be associated with a clearly defined taxon should also be deposited in recognised international databases that adhere to FAIR principles, such as BOLD (Barcode of Life Data Systems) or GenBank.

In addition, the data on species occurrences may be added to the database of the Global Biodiversity Information Facility (GBIF)⁸. Associated metadata, such as locations for type specimens or photographs, are being archived according to the requirements of the data repositories.

Taxonomic information relating to sequences may also be archived in custom databases with links to the sequences and organisms associated with them. The custom databases may be hosted on the BIOPOLIS project website⁹. However, it's crucial to consider FAIR principles when developing these custom databases. This includes ensuring proper documentation, open access options, and data formats compatible with established standards for long-term

⁷ BOLD: <http://www.barcodinglife.org/>

⁸ Global Biodiversity Information Facility (GBIF): <http://www.gbif.org>

⁹ BIOPOLIS project website: <https://www.biopolis.pt/>

usability. A noteworthy example is MIMt (Mass Identification in Metagenomic tests), a curated 16S rRNA database specifically designed for rapid species identification. Developed and maintained by BIOPOLIS members, MIMt is publicly available at <https://mimt.bu.biopolis.pt> which makes data available through well-known data formats.

3.2.3. Molecular Biological or Field Methods

To align with the Open Science principles of the Horizon Europe program, this Data Management Plan (DMP) prioritises transparency and reproducibility of research methods. To achieve this, both molecular biology protocols and field collection procedures should be archived using two complementary approaches:

- i) Publication of the methodologies or protocols in
 - a) Peer-reviewed Open Access publications: The aim is to publish in-depth descriptions of the methodologies in peer-reviewed Open Access journals. This allows for scientific scrutiny and fosters adoption by the wider research community.
 - b) Non-peer-reviewed public archives: Preprints of the methodologies will also be deposited in public archives like bioRxiv. This offers rapid dissemination and enables early feedback from colleagues, potentially accelerating refinement and replication efforts. Alternatively, protocols may be submitted to *protocols.io*¹⁰, a secure platform specifically designed for developing and sharing reproducible methods. *Protocols.io* facilitates clear and well-structured documentation of methodologies, version control, and community discussion, promoting transparency and replicability. While it is not a peer-reviewed platform, the community discussion feature can provide valuable feedback and contribute to overall method quality.

¹⁰ Protocols.io: <https://protocols.io>

ii) BioStudies database: Additionally, transcripts or videos documenting the methods will be made available on the project page within the BioStudies database. This online resource facilitates data discovery and fosters collaboration by providing easily accessible descriptions of protocols.

These two approaches foster transparency and adherence to the ORD principles. They ensure the long-term accessibility and potential reuse of valuable research methodologies.

3.2.4. Ecological Results Dissemination

BIOPOLIS is firmly committed to promoting open science practices, aligning with the Horizon Europe Open Science Policy (previously Horizon 2020 Open Research Data principles). This commitment includes ensuring the accessibility and reusability of ecological data generated at BIOPOLIS.

As part of this commitment, BIOPOLIS utilises BioStudies, a well-established open-access database platform, for archiving the primary ecological datasets within its projects. This open-access database facilitates data discovery, transparency, and responsible data sharing and serves as the chosen online platform for archiving the primary ecological datasets within BIOPOLIS projects.

For projects that involve gridded or time series environmental data (e.g., oceanographic data, atmospheric data, remote sensing imagery, etc), ERDDAP can be a valuable complementary resource. ERDDAP (the Environmental Research Division's Data Access Program) is a data server that provides a simple and consistent way to download subsets of scientific datasets in common file formats and make graphs and maps. There are numerous ERDDAP servers hosted by various institutions and organisations worldwide. BIOPOLIS recognizes the value of ERDDAP for facilitating data exploration and access and has set up one ERDDAP server at BIOPOLIS (<https://erddap.at.biopolis.pt/>) to share and distribute data.

BioStudies' commitment to open access, detailed metadata, robust search functionalities, and version control aligns with the FAIR (Findable, Accessible, Interoperable, Reusable) data

principles. This, along with potential integration with ERDDAP's functionalities, ensures the ecological data generated by BIOPOLIS projects is discoverable, accessible, and reusable for the benefit of the broader scientific community.

In addition to these repositories, ecological data collected in the scope of LTSE (Long-Term Socio-Ecological Research) Sabor, managed by BIOPOLIS, has been stored in the Dynamic Ecological Information Management System - Site and dataset registry¹¹ (DEIMS-SDR). This repository follows the guidelines of eLTER RI.

3.2.5. Camera Trap Images

BIOPOLIS is committed to archiving camera trap images according to FAIR principles (Findable, Accessible, Interoperable, Reusable). We are considering established platforms like eMammal¹² and Wildlife Insights¹³ for camera trap image storage. Each platform's advantages and limitations are being evaluated to select the most suitable option. The final evaluation and recommendations will be made by the DM&OAC. Utilising multiple platforms might be considered to maximise data dissemination.

Camera trap data, especially images, require substantial storage capacity. To address this it is recommended a combined approach:

- i) In-situ Storage: for initial processing and organisation. This temporary storage ensures data security and facilitates initial work on the data.
- ii) Cloud-based Archiving with Platform Integration: The DM&OAC will recommend a suitable platform (mentioned above) for long-term archiving and data management. This integration ensures efficient data organisation, accessibility, and potential access to analysis tools offered by the chosen platform. Importantly, it also minimises issues arising from limited local storage capacity.

¹¹ Dynamic Ecological Information Management System (DEIMS-SDR): <https://deims.org/>

¹² eMammal: <https://emammal.si.edu>

¹³ Wildlife Insights: <https://www.wildlifeinsights.org>

3.2.6. GPS-tagged Animal Data

BIOPOLIS recommends utilising Movebank Data Repository¹⁴ as the primary platform for archiving and managing animal movement data collected by GPS tags. Movebank is a well-established repository specifically designed for animal tracking data, offering several advantages:

- Alignment with FAIR Principles: Movebank strives to uphold the FAIR (Findable, Accessible, Interoperable, Reusable) data principles.
- Endorsed by Leading Journals and Publishers: Movebank's commitment to FAIR data practices makes it a recommended repository by several journals and publishers, including the British Ecological Society, the Journal of Zoology, Oikos, and Springer Nature. This endorsement simplifies compliance with data archiving policies for publications in these outlets.

By prioritising Movebank for archiving GPS-tagged animal data, BIOPOLIS contributes to a centralised repository specifically designed for this type of data. This centralised approach facilitates discovery, promotes standardised data management practices, and ultimately advances research in animal ecology and movement patterns.

3.2.7. Vegetation Data

Data on vegetation cover, species abundance, height, DBH (diameter at breast height), and phenological stage should be deposited in established ecological data repositories that align with FAIR principles. Examples include the Global Biodiversity Information Facility (GBIF), the Global Index of Vegetation Databases (GIVD)¹⁵, and the System for Information on Variety and Management of Algerian Forests (SiVIM)¹⁶. Establishing permanent forest vegetation plots will also be explored and data will be made available through <https://www.forestplots.net/pt>.

¹⁴ Movebank platform: <https://www.movebank.org>

¹⁵ Global Index of Vegetation Databases (GIVD): <https://www.givd.info/>

¹⁶ System for Information on Variety and Management of Algerian Forests (SiVIM): <http://www.sivim.info/sivi/>

To ensure compliance with FAIR principles, it is recommended to archive data (spectral vegetation indices, biophysical parameters, vegetation cover, productivity, phenology, seasonality, and spatial heterogeneity) in a recognized repository like the BioStudies database¹⁷. This resource is specifically designed for biological data and promotes data findability and interoperability. Alternatively, Zenodo¹⁸ may be considered for data archiving, but only if BioStudies is not deemed suitable after careful evaluation by the DM&OAC.

3.2.8. Land Use and Land Cover Data

There are ongoing efforts by the Food and Agricultural Organisation of the United Nations (FAO)¹⁹ regarding Land Use and Land Cover (LULC) data, and the organisation and harmonisation of this data in the databases (e.g, the Global Land Cover-SHARE (GLC-SHARE) database²⁰). LULC data generated within the scope of BIOPOLIS gathered via e.g. aerial photographs, GPS and Geographic Information System (GIS) and Remote sensing (RS) may be archived in the above-mentioned database. An assessment of the suitability of other platforms and archives will be performed by the DM&OAC.

3.2.9. Conventional Biodiversity Survey Data

This DMP adheres to the Darwin Core Standard (DwC)²¹ for archiving biodiversity data, including information on species presence/absence, abundance, and other relevant ecological data. The DwC provides a robust, yet adaptable framework for compiling biodiversity information from diverse sources, ensuring data consistency and interoperability. Sharing data via recognised repositories like GBIF maximises data findability and accessibility for the scientific community, fostering data reuse and promoting biodiversity research advancements.

¹⁷ Biostudies: <https://www.ebi.ac.uk/biostudies/>

¹⁸ Zenodo: <https://zenodo.org/>

¹⁹ Food and Agricultural Organisation of the United Nations (FAO): <http://fao.org>

²⁰ Global Land Cover - SHARE (GLC-SHARE): <http://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1036355/>

²¹ Darwin Core Standard (DwC): <https://www.gbif.org/darwin-core>

3.3. Data Analysis Methods

Computational methods used for data processing, statistical analysis, and visualisation should be archived in two ways:

- 1) Published as Supplementary Material: Scripts will be included as supplementary files for research papers published in Open Access journals.
- 2) BIOPOLIS GitHub Repository: Whenever possible, scripts should be deposited on the newly created BIOPOLIS GitHub page (<https://github.com/cibio-biopolis>), following details outlined in the SOFTWARE section.

3.4. Data Dissemination - Sharing and Publishing

BIOPOLIS is committed to promoting Open Science practices. This commitment aligns with the principles of the Horizon Europe program, particularly the FAIR (Findable, Accessible, Interoperable, Reusable) data management principles. to promote open science practices by adhering to the FAIR principles (Findable, Accessible, Interoperable, Reusable). This ensures that the data generated by BIOPOLIS projects is discoverable, accessible, and usable by other researchers.

All BIOPOLIS project data (raw and processed) should be deposited in the BioStudies database. BioStudies offers a flexible platform for archiving biological research data, including descriptions of studies; links to data stored elsewhere (e.g., GenBank, ENA, BOLD); supplementary data not suitable for structured archives (specific tables, protocols, photos, videos); and links to publications associated with the data.

Each BIOPOLIS project webpage should link directly to the corresponding BioStudies entry, providing easy access to all relevant data and metadata. Project-specific data may also be archived in specialised repositories for camera trapping (e.g., eMammal, Agouti, Wildlife Insights), GPS-tagged animals (e.g., Movebank), and vegetation plots/remotely sensed data (as described elsewhere).

All data collected or produced by BIOPOLIS should be, when possible, openly available for public reuse under a Creative Commons (CC) licence. By default, the CC Attribution 4.0

International (CC BY 4.0²²) licence should be used which allows others to i) Redistribute the material in any format or medium, and ii) Build upon and adapt the data for any purpose, including commercial. However, users must provide appropriate attribution (crediting BIOPOLIS/CIBIO as the source), link to the licence, and indicate if changes were made.

3.5. Software

BIOPOLIS is committed to open science practices. BIOPOLIS is committed to open science, emphasising transparency and accessibility of research methods and data to allow other researchers to understand the findings and verify and potentially replicate them. GitHub, a version control system originally designed for software development, has become a powerful tool for open science that may be used to share code, data, and protocols facilitating transparency and reproducibility. By fostering collaboration and scrutiny, open science practices facilitated by GitHub help to ensure the reliability and robustness of scientific discoveries. Therefore, all software and routines developed during the project should be publicly available on GitHub preferentially under <https://github.com/cibio-biopolis>. To ensure usability and reproducibility, software should adhere to well-established standards for:

- Packaging: Software will be packaged in a way that facilitates easy distribution and installation (e.g., Conda).
- Versioning: We will implement a clear version control system to track changes and ensure stability.
- Documentation: Comprehensive documentation will be provided to guide users in understanding and using the software.
- Updating: Mechanisms will be established for keeping the software up-to-date and addressing any issues.
- Installation: Straightforward instructions will be provided for users to install the software.

²² <https://creativecommons.org/licenses/by/4.0/>

The source code should be freely available under permissive open-source licences such as GNU General Public License v3 (GNU GPLv3), Apache License (Version 2.0) or MIT License. These permissive licences grant users various freedoms, such as allowing them to use, modify, and distribute the software.

3.6. DMP Implementation

As detailed in the previous sections of this DMP, all data and analysis methods should be archived in well-established public databases thus ensuring long-term preservation of the data. BIOPOLIS prioritises archiving data and analysis methods in well-established public databases. Many of these repositories are certified and part of the ELIXIR core infrastructure. Certification signifies adherence to high standards for data preservation and longevity. Utilising these platforms ensures data is: Findable (Discoverable by researchers through search engines and other tools); Accessible (Freely available for download and analysis without excessive restrictions); and Interoperable (Uses standardised formats and protocols facilitating data exchange and integration with other datasets).

It is paramount that researchers have expertise in FAIR data principles and utilising chosen archive platforms. To address this, targeted training workshops will be organised when necessary to cover relevant topics to equip researchers with the knowledge and skills for effective data management and archiving.

3.7. Data Security, Research Ethics, and Animal Welfare

As described in the previous sections of this DMP, data security is ensured by archiving data in established public repositories. These repositories implement robust security measures to protect data integrity and confidentiality. Additionally, the Data Management and Open Access Committee (DM&OAC) will monitor data access and usage.

In connection with this and the scope of the BIOPOLIS Knowledge Management System - KMS (D7.4a), a multilevel access permission system will be installed and implemented to reduce the possibility of wrong uses of the data, as well as protecting the interest of the partners.

BIOPOLIS adheres to the highest ethical standards and ensures full compliance with the Nagoya Protocol on Access and Benefit Sharing (ABS), EU Regulation No 511/2014, and Portuguese Decree-Law 1222/2017. These regulations govern the collection and analysis of genetic samples and the subsequent generation of BIOPOLIS project data.

The research activities of BIOPOLIS are monitored by the Ethics and Animal Welfare Committee to guarantee that all research is conducted following the best practices and the highest international standards. As outlined in the *Implementation Plan for Ethics and Animal Welfare Guidelines (Deliverable 4.5)*, the Ethics and Animal Welfare Committee benefits from the experience and diverse perspectives University of Montpellier's research and academic community. Their collaboration contributes to the development of strong and comprehensive BIOPOLIS guidelines for ethical research practices.

4. DMP Implementation Actions and Status

4.1. DMP Implementation Status

BIOPOLIS is committed to open data practices and leverages various public archives to ensure data accessibility. This commitment aligns with the Horizon Europe program's emphasis on FAIR data management principles.

Leveraging the expertise of BIOPOLIS members who are responsible for PORBIOTA and coordinate the Portuguese LifeWatch node, BIOPOLIS benefits from a strong foundation in metadata publication²³. The DMP outlines the specific data repositories chosen for different data types. Here are some examples:

²³ LifeWatch:

<https://metadatalogue.lifewatch.eu/srv/eng/catalog.search#/search?any=portugal>

- Sequence Data: Raw sequence reads and associated sample metadata are deposited in the European Nucleotide Archive (ENA) (e.g., accession numbers [ERX433270](#), [ERX4332052](#), or [SRR10012545](#)). Processed sequence data, such as assembled genomes (e.g., [GCA_903995535.1](#) and [GCA_903995505.1](#)), are also archived.
- DNA Barcodes: Significant effort has been dedicated to sharing data through the Barcode of Life Database (BOLD)²⁴, an open-access repository for DNA barcodes used in species identification. Over 10,000 specimens from more than 3,500 species, primarily collected in Portugal, Spain, and African countries (e.g., Sao Tome e Principe, Guinea-Bissau), have been published on BOLD.
- Taxonomic Data: BIOPOLIS actively contributes to the Global Biodiversity Information Facility (GBIF) platform. Over 4,000 species observations have been published on GBIF to share taxonomic data with the broader scientific community (see <https://www.gbif.org/publisher/37560003-ae75-43c3-95aa-48c98797aa2a>).
- Long-Term Ecological Data: Information on LTSEr Sabor as well as 33 ecological datasets have already been published in DEIMS-SDR (see <https://deims.org/45722713-80e3-4387-a47b-82c97a6ef62b>).
- BioStudies: This archive is another key platform for BIOPOLIS data sharing. Tens of datasets from BIOPOLIS have already been published on BioStudies (see e.g., [S-BSST920](#) and [S-EPMC9259742](#)).

While there is no direct cost for data storage in these repositories, ensuring FAIR compliance requires investment from researchers and technicians. Time spent on data upload, curation, and metadata creation translates to a cost that BIOPOLIS aims to minimise. Therefore, to achieve this, BIOPOLIS is considering the development of clear internal procedures for semi-automatic data FAIRification. This approach will streamline the process, reduce time commitment, optimise resource allocation, and increase the probability of adhesion to make data FAIR.

²⁴ BOLD: <http://www.barcodinglife.org/>

4.2. DMP Implementation Actions

To ensure the long-term accessibility and usability of our research data and further boost the implementation of the DMP, BIOPOLIS recently established the Data Management and Open Access Committee (DM&OAC). This dedicated committee will play a crucial role in overseeing all aspects of data management and open access within BIOPOLIS. Their responsibilities include:

- overall validation of Data Management Plans (DMPs) submitted by individual projects.
- planning, implementing, and monitoring data management policies and procedures, ensuring alignment with open access principles.
- reviewing and updating DMPs according to established timelines outlined in the project Grant Agreement (GA).

The following actions, and others to be defined by the DM&OAC, will be taken to achieve it:

- **Dissemination:** The DMP will be disseminated to all project partners and researchers involved in BIOPOLIS activities. Training sessions will be conducted to ensure understanding and proper implementation of the outlined procedures.
- **Integration with Research Processes:** Data management practices will be integrated into the research workflow from project inception. This ensures data is collected, documented, and archived according to DMP guidelines.
- **Version Control and Updates:** The DMP will be a living document. The BIOPOLIS Data Management and Open Access Committee (DM&OAC) will be responsible for reviewing and updating the DMP as needed, reflecting changes in research practices, data types, or archiving technologies.
- **Data Security.** As described in the DMP, data security is ensured by archiving data in established public repositories. These repositories implement robust security measures to protect data integrity and confidentiality. Additionally, the Data Management and Open Access Committee (DM&OAC) will monitor data access and usage. In connection with this and the scope of the BIOPOLIS Knowledge Management System - KMS (D7.4a), a multilevel access permission system will be installed and

implemented to reduce the possibility of wrong uses of the data, as well as protecting the interest of the partners.

Finally, it is paramount that researchers have expertise in FAIR (Findable, Accessible, Interoperable, Reusable) data principles and utilise chosen archive platforms. To address this, targeted training workshops will be organised when necessary to cover relevant topics to equip researchers with the knowledge and skills for effective data management and archiving.

To handle the growing demands of big data, particularly from omics platforms, BIOPOLIS is actively investing in upgrading its computational infrastructure. This comprehensive upgrade will encompass several key areas, namely:

- **Storage Area Network (SAN) Expansion:** The SAN will be enhanced to provide the capacity and performance required for efficient data storage and retrieval.
- **Compute Node Upgrade:** More compute nodes will be deployed to expedite data processing tasks, particularly for complex analyses.
- **Large Scratch Storage Acquisition:** A dedicated high-performance scratch storage system will be acquired to facilitate the temporary storage and analysis of large datasets. This will significantly improve processing efficiency for big data analysis.
- **Network Enhancements:** Both internal network infrastructure and external bandwidth will be upgraded. The internal network will be optimised for efficient data movement within BIOPOLIS. Additionally, a tenfold increase in external network bandwidth is planned through the RCTS (Science, Technology, and Society Network). This will significantly improve data transfer speeds, facilitating communication and collaboration with external resources.

4.3. DMP Implementation KPIs

To monitor the effectiveness of the DMP, the following Key Performance Indicators (KPIs) will be tracked:

- **Data Management Training:** Number of researchers trained on data management best practices and FAIR principles.

- FAIR Data Archiving: Number of datasets deposited in FAIR-compliant repositories within the designated time frame.
- Metadata Completion: Completion rate of metadata associated with archived data, ensuring discoverability and usability.
- DMP Maintenance: Frequency of DMP review and update by the DM&OAC, reflecting evolving research practices and data types.
- Open Access Publications: Number of research papers published as open access, promoting knowledge dissemination.
- Data Reuse: Number of datasets published in BioStudies and other relevant repositories, fostering data reuse by the scientific community.
- Software Sharing: the number of software tools made available through platforms like GitHub, encouraging collaboration and innovation.

Regular monitoring of these KPIs will allow the DM&OAC to identify areas for improvement and ensure the DMP remains effective in supporting responsible data management practices throughout the BIOPOLIS project lifecycle

5. Concluding Remarks

The BIOPOLIS is committed to promoting open science practices within the research areas it encompasses, namely: Environment and Biodiversity assessment and Monitoring, Ecosystem Function and Restoration, Agrobiodiversity, Conservation and Competitiveness of local Genetic Resources and Farming systems. This commitment aligns with the Horizon Europe Open Science Policy program (previously Horizon 2020 Open Research Data principles).

We achieve this by adhering to the following key principles:

- Balancing Openness and Protection: BIOPOLIS prioritises responsible data management, balancing openness with data protection. We follow the principle of "as open as possible, as closed as necessary" to ensure research data accessibility while respecting legitimate restrictions.

- FAIR Data Management: To further promote open science, BIOPOLIS adheres to the FAIR data principles (Findable, Accessible, Interoperable, Reusable). This commitment aligns with the emphasis on open data within the Horizon Europe program and ensures that data generated by BIOPOLIS projects is:
 - Discoverable: Easily found by researchers through search engines and data repositories.
 - Accessible: Available to researchers with minimal restrictions for legitimate research purposes.
 - Interoperable: Usable with different software and data analysis tools.
 - Reusable: Well-documented and formatted to facilitate reuse in future research.
- Open Research Data (ORD): BIOPOLIS aims to implement a comprehensive ORD approach to guarantee data availability beyond the lifespan of individual projects. This ensures long-term data preservation and accessibility for the broader scientific community.

While this DMP carefully outlines current data management strategies, the exact types of data generated may evolve due to changes in BIOPOLIS research fields. The BIOPOLIS Data Management and Open Access Committee (DM&OAC) will proactively monitor these changes and update this DMP accordingly. This includes adapting to evolving data archiving practices within the wider research community.

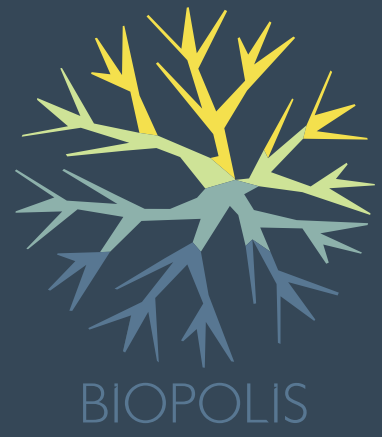
The selection of public databases for data archiving offers several advantages:

- Compliance with FAIR principles: By exploring the functionalities of these databases, we can ensure the discoverability, accessibility, interoperability, and reusability of the data.
- Sustainable Data Availability: Linking data and metadata (as described in the Data Dissemination section) within these platforms promotes long-term accessibility beyond the project duration.



- Accommodation of New Data Types: The flexibility of many chosen databases minimises the need for significant adjustments when incorporating new data types in the future.

The BIOPOLIS recognises the importance of responsible data stewardship. By adhering to the principles outlined in this DMP, we aim to maximise the value and impact of our research for the scientific community and society at large.



This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Grant Agreement Number 857251.